A Reproducing Kernel Hilbert Space Framework for Information-Theoretic Learning

Jian-Wu Xu, *Student Member, IEEE*, António R. C. Paiva, *Student Member, IEEE*, Il Park (Memming), and Jose C. Principe, *Fellow, IEEE*

***Abstract*—This paper provides a functional analysis perspective of information-theoretic learning (ITL) by defining bottom-up a reproducing kernel Hilbert space (RKHS) uniquely determined by the symmetric nonnegative definite kernel function known as the cross-information potential (CIP). The CIP as an integral of the product of two probability density functions characterizes similarity between two stochastic functions. We prove the existence of a one-to-one congruence mapping between the ITL RKHS and the Hilbert space spanned by square integrable probability density functions. Therefore, all the statistical descriptors in the original information-theoretic learning formulation can be rewritten as algebraic computations on deterministic functional vectors in the ITL RKHS, instead of limiting the functional view to the estimators as is commonly done in kernel methods. A connection between the ITL RKHS and kernel approaches interested in quantifying the statistics of the projected data is also established.**

***Index Terms*—Cross-information potential, information-theoretic learning (ITL), kernel function, probability density function, reproducing kernel Hilbert space (RKHS).**

I. Introduction

A reproducing kernel Hilbert space (RKHS) is a special Hilbert space associated with a kernel such that it reproduces (via an inner product) each function in the space or, equivalently, a space where every point evaluation functional is bounded. Let $\mathcal{H}$ be a Hilbert space of real-valued functions on a set E, equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a real-valued bivariate function $\kappa(x, y)$ on $E \times E$. Then the function $\kappa(x, y)$ is said to be nonnegative definite if, for any finite point set $\{x_1, x_2, \ldots, x_n\} \subseteq E$ and for any not all zero corresponding real numbers $\{\alpha_1, \alpha_2, \ldots, \alpha_n\} \subseteq \mathbb{R}$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \kappa(x_i, x_j) \geq 0.$$

Any nonnegative definite bivariate function $\kappa(x, y)$ is a reproducing kernel because of the following fundamental theorem.

Manuscript received September 25, 2007; revised July 15, 2008. First published August 29, 2008; current version published November 19, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Matz. This work was supported in part by the National Science Foundation under Grant ECS-0601271.

The authors are with the Computational NeuroEngineering Laboratory, Electrical and Computer Engineering Department, University of Florida, Gainesville, FL 32611 USA (e-mail: jianwu@cnel.ufl.edu; arpaiva@cnel.ufl.edu; memming@cnel.ufl.edu; principe@cnel.ufl.edu).

Digital Object Identifier 10.1109/TSP.2008.2005085

Theorem 1 (Moore–Aronszajn): Given any nonnegative definite function $\kappa(x, y)$, there exists a uniquely determined (possibly infinite dimensional) Hilbert space $\mathcal{H}$ consisting of functions on E such that

$$\text{(I)} \quad \forall x \in E,\ \kappa(x, \cdot) \in \mathcal{H}$$
$$\text{(II)} \quad \forall x \in E,\ \forall f \in \mathcal{H}, f(x) = \langle f, \kappa(x, \cdot) \rangle_{\mathcal{H}}.$$

Then $\mathcal{H} := \mathcal{H}_\kappa$ is said to be a reproducing kernel Hilbert space with reproducing kernel κ. Property II is called the *reproducing property* of $\kappa(x, y)$ in $\mathcal{H}_\kappa$.

The existence of a reproducing kernel Hilbert space corresponding to any symmetric and nonnegative definite kernel function is one of the most fundamental results [1]. The reproducing kernel Hilbert space framework was instrumental to providing practical solutions to a class of differential equations (Green's functions). But it was not until 1943 that Aronszajn systematically developed the general theory of RKHS and coined the term "reproducing kernel" [2].

The RKHS framework also provides a natural link between stochastic process and deterministic functional analysis. It was Parzen who first introduced the RKHS methodology in statistical signal-processing and time-series analysis in the late-1950s. The essential idea is that there exists a congruence map between the RKHS of random variables spanned by the random process and its covariance function $R(t, s) = E[\mathrm{X(t)X(s)}]$, which determines a unique RKHS, denoted as $\mathcal{H}_R$. Note that the kernel includes the second-order statistics of the data through the expected value (a data-dependent kernel) and Parzen clearly illustrated that the RKHS offers an elegant functional analysis framework for minimum variance unbiased estimation of regression coefficients, least squares estimation of random variables, detection of signals in Gaussian noise, and others [3]–[5]. In the early 1970s, Kailath *et al.* presented a series of detailed papers on the RKHS approach to detection and estimation problems to demonstrate its superiority in computing likelihood ratios, testing for nonsingularity, bounding signal detectability, and determining detection stability [6]–[10]. RKHS concepts have also been extensively applied to a wide variety of problems in optimal approximation including interpolation and smoothing by spline functions in one or more dimensions (curve and surface fitting) [11]. De Figueiredo took a different approach to apply RKHS in nonlinear system and signal analysis [12]. He built the RKHS bottom-up using arbitrarily weighted Fock spaces that played an important role in quantum mechanics [13]. The spaces are composed of polynomials or power series in either scalar or multidimensional variables. The generalized Fock spaces have

been also applied to nonlinear system approximation, semiconductor device characteristics modeling, and neural networks [12].

Recent work on support vector machines by Vapnik rekindled the interest in RKHS for pattern recognition [14], [15] but with a different twist. Here RKHS is used primarily as a high-dimensional feature space where the inner product is efficiently computed by means of the kernel trick. A nonnegative definite kernel function κ (e.g., Gaussian, Laplacian, polynomial, and others [16]) nonlinearly projects the data sample-by-sample into a high-dimensional RKHS, denoted as $\mathcal{H}_\kappa$, induced by κ. For separability (Cover theorem [17]) it is advantageous to consider the learning problem in the RKHS because of its high dimension. When learning algorithms can be expressed in terms of inner products, this nonlinear mapping becomes particularly interesting and useful since kernel evaluations in the input space *implicitly* compute inner products of the transformed data in the RKHS without *explicitly* using or even knowing the nonlinear mapping. Therefore, exploiting the linear structure of the RKHS, one can elegantly build a nonlinear version of a linear algorithm based on inner products and short circuit the need for iterative training methods as necessary in artificial neural networks. Essentially a kernel method is a shallow (one layer) neural network whose parameters can be analytically computed given the training set data. Numerous kernel-based learning algorithms have been proposed in machine learning [18]. For example, kernel principal component analysis [19] and kernel independent component analysis [20] are some of the most well-known kernel-based learning methods. However, this RKHS structure is given by the kernel $\mathcal{H}_\kappa$ and is therefore data independent, unlike $\mathcal{H}_R$.

More recently, a research topic called information-theoretic learning (ITL) has emerged independently [21], [22]. ITL is an optimal signal-processing technique that combines information and adaptive filtering theories to implement new cost functions for adaptation that lead to "information filtering" [22] without requiring a model of the data distributions. ITL builds upon the concepts of Parzen windowing applied to Rényi's entropy [23] to obtain a sample by sample estimator of entropy directly from pairs of sample interactions. By utilizing the quadratic Rényi's measure of entropy and approximations to the Kullback–Leibler divergence, ITL proposes alternate similarity metrics that quantify higher order statistics directly from the data in a nonparametric manner. ITL has achieved excellent results on a number of learning scenarios such as clustering [24], [25], feature extraction [26], function approximation [27], and blind equalization [28]. The centerpiece in all these engineering applications is the *information potential* (IP) estimator, which estimates the argument of the logarithm of Rényi's quadratic entropy using Parzen window kernels. However, its direct relation to entropy [and therefore to the probability density function (pdf) of the data] and the Cauchy–Schwarz divergence between pdfs evaluated in RKHS [29] indicates that it may be alternatively formulated as a data-dependent kernel transformation that transfers the statistical properties of the data to a different RKHS as Parzen's $\mathcal{H}_R$ does.

The main focus of this paper is exactly to construct an RKHS framework for information-theoretic learning (ITL RKHS). The ITL RKHS is directly defined on a space of pdfs with a kernel defined by inner products of pdfs. Since all the statistical information of the input is represented in its pdf, the ITL RKHS will be *data dependent* and is suitable for directly obtaining statistics from the data using inner products. Moreover, it still takes full advantage of the kernel trick to evaluate IP directly from data and the other descriptors of ITL.

The remainder of this paper is organized as follows. In Section II, we briefly introduce the concept of information-theoretic learning and relevant statistical descriptors. We then propose the ITL RKHS framework in Section III. The new understanding and the ITL descriptors are rewritten in the RKHS framework in Section IV. A theoretical lower bound for the information potential is proved in Section V. We further connect ITL RKHS to the statistics in kernel space in Section VI. We discuss some specific issues in Section VII and conclude in Section VIII.

II. Information-Theoretic Learning

The initial goal of ITL was to propose alternative cost functions for adaptive filtering that would quantify higher order moments of the error sequence [27]. Since entropy characterizes the uncertainty in the error, it was judged a good candidate. The difficulty of Shannon's entropy resides on nonparametric estimation within the constraints of optimal filtering (e.g., smooth costs). For this reason, we embraced a generalization of Shannon's entropy proposed by Rényi [22]. Rényi's entropy definition is based on the so called quasi-linear mean, which is the most general mean compatible with Kolmogorov's axiomatics [30]. It is a parametric family of entropies given by

$$H_\alpha(\mathrm{X}) = \frac{1}{1-\alpha}\log\int f^\alpha(x)\mathrm{d}x, \quad \alpha > 0$$

where $f(x)$ is the pdf of the continuous random variable X. It is easy to show that the limit of $\alpha \to 1$ yields Shannon's entropy and [23] shows that the singularity is not essential. Perhaps a more insightful interpretation is to observe

$$\int f^\alpha(x)\mathrm{d}x = E\left[f^{\alpha-1}(x)\right]$$

i.e., the argument of the logarithm of Rényi's entropy is the α-*order moments of the pdf* of the data.

The initial interest of Rényi's definition for ITL was the existence of a practical nonparametric estimator for the quadratic Rényi's entropy $(\alpha = 2)$ defined as

$$H_2(\mathrm{X}) = -\log\int f^2(x)\mathrm{d}x = -\log E\left[f(x)\right] \tag{1}$$

i.e., the $-\log$ of the first moment of the pdf. Since the logarithm function is monotonic, the quantity of interest in adaptive filtering is the first moment of the pdf itself

$$\mathcal{V}(\mathrm{X}) = \int f^2(x)\mathrm{d}x \tag{2}$$

which is called the *information potential,*[1] so named due to a similarity with a potential field in physics [21].

A nonparametric asymptotically unbiased and consistent estimator for a given pdf $f(x)$ is defined as [31]

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\kappa(x, x_i) \tag{3}$$

where $\kappa(\cdot,\cdot)$ is called the Parzen window, or kernel. Here the Parzen kernel will be chosen as a symmetric nonnegative definite function just like in kernel-based learning theory, such as the Gaussian, polynomial, etc. [16]. Then by evaluating the expectation of Parzen's pdf approximation in (2), the integral can be directly estimated from the data as

$$\hat{V}(\mathrm{X}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x_i, x_j) \tag{4}$$

where $\{x_i\}_{i=1}^{N}$ is the data sample and N is the total number, which is the estimator for IP. The concept and properties of information potential (and its estimator) have been mathematically studied and a new criterion based on information potential proposed, called the minimization error entropy (MEE), to adapt linear and nonlinear systems [27]. MEE serves as an alternative cost function to the conventional mean square error (MSE) in linear/nonlinear filtering with several advantages in performance when the error distribution is not Gaussian. If we think for a moment, we see the big difference between MSE and MEE: MSE is the *second-order moment of the data* and MEE is the *first moment of the pdf of the data*. Since all the information contained in the random variable is represented in its pdf, we can expect better performance from the latter than from MSE.

In information theory, mutual information is used to quantify the divergence between the joint pdf and the product of marginal pdfs of two random variables. Another well-known divergence measure is the Kullback–Leibler divergence [32]. However, both are difficult to estimate in practice without imposing simplifying assumptions about the data. Numerical methods are required to evaluate the integrals. This IP and two divergence measures among pdfs, one based on their Euclidean distance and the other on Cauchy–Schwarz inequality, have been proposed to surpass these limitations [21].

Given two probability density functions $f(x)$ and $g(x)$, their Euclidean divergence is defined as

$$\begin{aligned} D_{\mathrm{ED}}(f,g) &= \int \left(f(x) - g(x)\right)^2 \mathrm{d}x \\ &= \int f(x)^2 \mathrm{d}x - 2\int f(x)g(x)\mathrm{d}x \\ &\quad + \int g(x)^2 \mathrm{d}x. \end{aligned} \tag{5}$$

[1]Note that in previously published papers, we called the estimator for this quantity (4) as the information potential. In this paper, we generalize the concept and call the statistical descriptor behind it (2) as the IP and refer to (4) as the estimator of IP. The physical interpretation still holds.

The divergence measure based on Cauchy–Schwarz inequality is given by

$$D_{\mathrm{CS}}(f,g) = -\log \frac{\int f(x)g(x)\mathrm{d}x}{\sqrt{\left(\int f^2(x)\mathrm{d}x\right)\left(\int g^2(x)\mathrm{d}x\right)}}. \tag{6}$$

Notice that both $D_{\mathrm{ED}}(f,g)$ and $D_{\mathrm{CS}}(f,g)$ are greater than or equal to zero, and the equality holds if and only if $f(x) = g(x)$. Notice the form of the integrals. We have in both the first moment of each pdf and a new term $\int f(x)g(x)\mathrm{d}x$ that is the first moment of the pdf $g(x)$ over the other pdf $f(x)$ (or vice versa), which is called the *cross-information potential* (CIP) [21]. CIP measures the similarity between two pdfs as can be expected due to its resemblance to Bhattacharyya distance and other distances, as explained in [24]. CIP appears both in Euclidean and Cauchy–Schwarz divergence measures. If one substitutes $g(x)$ by $f(x)$ in CIP, it becomes the argument of Rényi's quadratic entropy. As expected, all these terms can be estimated directly from data as in (4).

The Euclidean (5) and Cauchy–Schwarz divergence (6) can be easily extended to two-dimensional random variables. As a special case, if we substitute the pdfs f and g in (5) and (6) by the joint pdf $f_{1,2}(x_1, x_2)$ and the product of marginal pdfs $f_1(x_1)f_2(x_2)$ for random variables $\mathrm{X}_1, \mathrm{X}_2$, respectively, we obtain the Euclidean quadratic mutual information as [21]

$$\begin{aligned} I_{\mathrm{ED}}(\mathrm{X}_1,\mathrm{X}_2) = &-2\iint f_{1,2}(x_1,x_2)f_1(x_1)f_2(x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &+\iint f_{1,2}^2(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 + \iint f_1^2(x_1)f_2^2(x_2)\mathrm{d}x_1\mathrm{d}x_2 \end{aligned} \tag{7}$$

and the Cauchy–Schwarz quadratic mutual information as [21]

$$\begin{aligned} &I_{\mathrm{CS}}(\mathrm{X}_1,\mathrm{X}_2) \\ &\quad = -\log \frac{\iint f_{1,2}(x_1,x_2)f_1(x_1)f_2(x_2)\mathrm{d}x_1\mathrm{d}x_2}{\sqrt{\iint f_{1,2}^2(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 \iint f_1^2(x_1)f_2^2(x_2)\mathrm{d}x_1\mathrm{d}x_2}}. \end{aligned} \tag{8}$$

If and only if the two random variables are statistically independent, then $I_{\mathrm{ED}}(\mathrm{X}_1,\mathrm{X}_2) = 0$ (and also $I_{\mathrm{CS}}(\mathrm{X}_1,\mathrm{X}_2) = 0$). The appeal of these four divergences is that every term in the equations can be directly estimated, yielding again practical algorithms. Hence, their manipulation can be used for unsupervised learning algorithms such as independent component analysis [33] and clustering [34].

The IP estimator (4) can be interpreted in an RKHS. Indeed, using the reproducing property of κ, any symmetric nonnegative definite kernel function can be written as $\kappa(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}_\kappa}$, where $\Phi(x)$ is the nonlinearly transformed data in the RKHS $\mathcal{H}_\kappa$ induced by the kernel

function and the inner product is performed in $\mathcal{H}_\kappa$. Therefore, we can rewrite (4) as

$$\hat{V}(\mathrm{x}) = \left\langle \frac{1}{N}\sum_{i=1}^{N}\Phi(x_i), \frac{1}{N}\sum_{j=1}^{N}\Phi(x_j)\right\rangle_{\mathcal{H}_\kappa} = \left\|\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i)\right\|^2.$$

Similar interpretations of the Cauchy–Schwarz divergence in $\mathcal{H}_\kappa$ were developed in [36]. As pointed out in the previous section, the RKHS $\mathcal{H}_\kappa$ is data independent since the kernel is predesigned and acts on individual data samples, which means that extra computation involving functional evaluations in $\mathcal{H}_\kappa$ is required when statistical quantities are estimated. The example of the IP estimator is still pretty simple and can exploit the kernel trick, but in general this may not be the case. The difficulty is that the inner product structure of $\mathcal{H}_\kappa$ is not translating the statistics of the data. This difficulty is not peculiar to ITL but extends to any of the kernel-based machine-learning algorithms. For instance, the estimator of kernel ICA can be estimated in $\mathcal{H}_\kappa$ but the covariance of the projected data is very hard to estimate [20]. Therefore, we conclude that it is unclear how to induce a general RKHS where the definition of inner products incorporating the statistical descriptors of the data will allow signal processing from basic principles. $\mathcal{H}_R$ is a step in this direction, but it only applies to Gaussian processes and is a linear mapping of the input data space, as will be discussed later.

III. RKHS Framework for ITL

From the various definitions in information-theoretic learning summarized above, we see that the most general quantity of interest is the integral of the product of two pdfs $\int f(x)g(x)\mathrm{d}x$, which we called the CIP. Therefore, this will be our starting point for the definition of the ITL RKHS that will include the statistics of the input data in the kernel.

A. The L_2 Space of PDFs

Let $\mathcal{E}$ be the set that consists of all square integrable one-dimensional probability density functions, i.e., $f_i(x) \in \mathcal{E}, \forall i \in \mathbb{I}$, where $\int f_i(x)^2\mathrm{d}x < \infty$ and $\mathbb{I}$ is an index set. We then form a linear manifold

$$\left\{\sum_{i\in I}\alpha_i f_i(x)\right\} \tag{9}$$

for any countable $I \subset \mathbb{I}$ and $\alpha_i \in \mathbb{R}$. Complete the set in (9) using the metric

$$\|f_i(x) - f_j(x)\| = \sqrt{\int (f_i(x) - f_j(x))^2\,\mathrm{d}x} \quad \forall i,j \in \mathbb{I} \tag{10}$$

and denote the set of all linear combinations of pdfs and its limit points by $L_2(\mathcal{E})$. $L_2(\mathcal{E})$ is an L_2 space on pdfs. Moreover, by the theory of quadratically integrable functions, we know that the linear space $L_2(\mathcal{E})$ forms a Hilbert space if an inner product is imposed accordingly. Given any two pdfs $f_i(x)$ and $f_j(x)$ in $\mathcal{E}$, we can define an inner product as

$$\langle f_i(x), f_j(x)\rangle_{L_2} = \int f_i(x)f_j(x)\mathrm{d}x \quad \forall i,j \in \mathbb{I}. \tag{11}$$

Notice that this inner product is exactly the CIP. This definition of inner product has a corresponding norm of (10). Hence, $L_2(\mathcal{E})$ equipped with the inner product (11) is a Hilbert space. However, it is not an RKHS because the inner product is not reproducing in $L_2(\mathcal{E})$, i.e., the point evaluation of any element in $L_2(\mathcal{E})$ cannot be represented via the inner product between two functionals in $L_2(\mathcal{E})$. Next we show that the inner product (11) is symmetric nonnegative definite, and by the Moore–Aronszajn theorem it uniquely defines an RKHS.

B. RKHS $\mathcal{H}_\mathcal{V}$ Based on $L_2(\mathcal{E})$

First, we define a bivariate function on the set $\mathcal{E}$ as

$$\mathcal{V}(f_i, f_j) = \int f_i(x)f_j(x)\mathrm{d}x \quad \forall i,j \in \mathbb{I}. \tag{12}$$

In RKHS theory, the kernel function is a measure of similarity between functionals. Notice that (12) corresponds to the definition of the inner product (11) and the cross-information potential between two pdfs; hence it is natural and meaningful to define the kernel function as $\mathcal{V}(f_i, f_j)$. Next, we show that (12) is symmetric nonnegative definite in $\mathcal{E}$.

Property 1 (Nonnegative Definiteness): The function (12) is symmetric nonnegative definite in $\mathcal{E}\times\mathcal{E} \longrightarrow \mathcal{R}$.

Proof: The symmetry is obvious. Given any positive integer N, any set of $\{f_1(x), f_2(x), \ldots, f_N(x)\} \subseteq \mathcal{E}$ and any not all zero real numbers $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$, by definition we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\mathcal{V}(f_i,f_j) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\int f_i(x)f_j(x)\mathrm{d}x \\ &= \int\left(\sum_{i=1}^{N}\alpha_i f_i(x)\right)\left(\sum_{j=1}^{N}\alpha_j f_j(x)\right)\mathrm{d}x \\ &= \int\left(\sum_{i=1}^{N}\alpha_i f_i(x)\right)^2\mathrm{d}x \geq 0.\end{aligned}$$

Hence, $\mathcal{V}(f_i, f_j)$ is symmetric nonnegative definite, and it is also a kernel function. □

According to the Moore–Aronszajn theorem, there is a unique RKHS, denoted by $\mathcal{H}_\mathcal{V}$, associated with the symmetric nonnegative definite function (12). We construct the RKHS $\mathcal{H}_\mathcal{V}$ bottom-up. Since the bivariate function (12) is symmetric and nonnegative definite, it also has an eigendecomposition by Mercer's theorem [35] as

$$\mathcal{V}(f_i, f_j) = \sum_{k=1}^{\infty}\lambda_k\psi_k(f_i)\psi_k(f_j) \tag{13}$$

where $\{\psi_k(f_i), k = 1, 2, \ldots\}$ and $\{\lambda_k, k = 1, 2, \ldots\}$ are sequences of eigenfunctions and corresponding eigenvalues of the kernel function $\mathcal{V}(f_i, f_j)$, respectively. The series above converges absolutely and uniformly on $\mathcal{E}\times\mathcal{E}$ [35].

Then we define a space $\mathcal{H}_{\mathcal{V}}$ consisting of all functionals $\mathcal{G}(\cdot)$ whose evaluation for any given pdf $f_i(x) \in \mathcal{E}$ is defined as

$$\mathcal{G}(f_i) = \sum_{k=1}^{\infty} \lambda_k a_k \psi_k(f_i) \tag{14}$$

where the sequence $\{a_k, k = 1, 2, \ldots\}$ satisfies the following condition:

$$\sum_{k=1}^{\infty} \lambda_k a_k^2 < \infty. \tag{15}$$

Furthermore, we define an inner product of two functionals in $\mathcal{H}_{\mathcal{V}}$ as

$$\langle \mathcal{G}, \mathcal{F} \rangle_{\mathcal{H}_{\mathcal{V}}} = \sum_{k=1}^{\infty} \lambda_k a_k b_k \tag{16}$$

where $\mathcal{G}$ and $\mathcal{F}$ are of form (14) and a_k and b_k satisfy (15).

It can be verified that the space $\mathcal{H}_{\mathcal{V}}$ equipped with the kernel function (12) is indeed a *reproducing kernel Hilbert space* and the kernel function $\mathcal{V}(f_i, \cdot)$ is a *reproducing kernel* because of the following two properties.

1) $\mathcal{V}(f_i, f_j)$ as a function of $f_i(x)$ belongs to $\mathcal{H}_{\mathcal{V}}$ for any given $f_j(x) \in \mathcal{E}$ because we can rewrite $\mathcal{V}(f_i, f_j)$ as

$$\mathcal{V}(f_i, \cdot)(f_j) = \sum_{k=1}^{\infty} \lambda_k b_k \psi_k(f_j), \quad b_k = \psi_k(f_i).$$

That is, the constants $\{b_k, k = 1, 2, \ldots\}$ become the eigenfunctions $\{\psi_k(f_i), k = 1, 2, \ldots\}$ in the definition of $\mathcal{G}$. Therefore

$$\mathcal{V}(f_i, \cdot) \in \mathcal{H}_{\mathcal{V}}, \quad \forall f_i(x) \in \mathcal{E}.$$

2) Given any $\mathcal{G} \in \mathcal{H}_{\mathcal{V}}$, the inner product between the reproducing kernel and $\mathcal{G}$ yields the function itself by the definition (16)

$$\begin{aligned} \langle \mathcal{G}, \mathcal{V}(f_i, \cdot) \rangle_{\mathcal{H}_{\mathcal{V}}} &= \sum_{k=1}^{\infty} \lambda_k a_k b_k \\ &= \sum_{k=1}^{\infty} \lambda_k a_k \psi_k(f_i) = \mathcal{G}(f_i). \end{aligned}$$

This is called the *reproducing property*.

Therefore, $\mathcal{H}_{\mathcal{V}}$ is an RKHS with the kernel function and inner product defined above. □

By the reproducing property, we can rewrite the kernel function (13) as

$$\begin{aligned} \mathcal{V}(f_i, f_j) &= \langle \mathcal{V}(f_i, \cdot), \mathcal{V}(f_j, \cdot) \rangle_{\mathcal{H}_{\mathcal{V}}} \\ \mathcal{V}(f_i, \cdot) &: f_i \mapsto \left[\sqrt{\lambda_k} \psi_k(f_i) \right], \quad k = 1, 2, \ldots. \end{aligned} \tag{17}$$

The reproducing kernel linearly maps the original pdf $f_i(x)$ into the RKHS $\mathcal{H}_{\mathcal{V}}$.

We emphasize here that the reproducing kernel $\mathcal{V}(f_i, f_j)$ is deterministic and data-dependent, by which we mean the norm of the transformed vector in the RKHS $\mathcal{H}_{\mathcal{V}}$ is dependent on the pdf of the original random variable because

$$\|\mathcal{V}(f_i, \cdot)\|^2 = \langle \mathcal{V}(f_i, \cdot), \mathcal{V}(f_i, \cdot) \rangle_{\mathcal{H}_{\mathcal{V}}} = \int f_i(x)^2 \mathrm{d}x.$$

This is very different from the reproducing kernel $\kappa(x, y)$ used in kernel-based learning theory. The norm of nonlinearly projected vector in the RKHS $\mathcal{H}_{\kappa}$ does not rely on the statistical information of the original data since

$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle_{\mathcal{H}_{\kappa}} = \kappa(0)$$

if we use translation-invariant kernel functions [16]. Moreover, if X is a random variable, $\Phi(\mathrm{X})$ is also a random variable in the RKHS $\mathcal{H}_{\kappa}$. The value of $\kappa(0)$ is a constant regardless of the original data. Consequently, the reproducing kernel Hilbert spaces $\mathcal{H}_{\mathcal{V}}$ and $\mathcal{H}_{\kappa}$ determined by $\mathcal{V}(f_i, f_j)$ and $\kappa(x, y)$, respectively, are very different in nature.

C. Congruence Map Between $\mathcal{H}_{\mathcal{V}}$ and $L_2(\mathcal{E})$

We have presented two Hilbert spaces: the Hilbert space $L_2(\mathcal{E})$ of pdfs and the RKHS $\mathcal{H}_{\mathcal{V}}$. Even though their elements are very different, there actually exists a one-to-one congruence mapping Ψ (isometric isomorphism) from RKHS $\mathcal{H}_{\mathcal{V}}$ onto $L_2(\mathcal{E})$ such that

$$\Psi\left(\mathcal{V}(f_i, \cdot)\right) = f_i. \tag{18}$$

Notice that the mapping Ψ preserves isometry between $\mathcal{H}_{\mathcal{V}}$ and $L_2(\mathcal{E})$ since by definitions of inner product (11) in $L_2(\mathcal{E})$ and (17) in $L_2(\mathcal{E})$

$$\begin{aligned} \langle \mathcal{V}(f_i, \cdot), \mathcal{V}(f_j, \cdot) \rangle_{\mathcal{H}_{\mathcal{V}}} &= \langle f_i(x), f_j(x) \rangle_{L_2} \\ &= \langle \Psi\left(\mathcal{V}(f_i, \cdot)\right), \Psi\left(\mathcal{V}(f_j, \cdot)\right) \rangle_{L_2}. \end{aligned}$$

That is, the mapping Ψ maintains the inner products in both $\mathcal{H}_{\mathcal{V}}$ and $L_2(\mathcal{E})$.

In order to obtain an explicit representation of Ψ, we define an orthogonal function sequence $\{\xi_m(x), m = 1, 2, \ldots\}$ satisfying

$$\int \xi_k(x) \xi_m(x) \mathrm{d}x = \begin{cases} 0, & k \neq m \\ \lambda_k, & k = m \end{cases}$$

and

$$\int \sum_{k=1}^{\infty} \psi_k(f_i) \xi_k(x) \mathrm{d}x = 1 \tag{19}$$

where $\{\lambda_k\}$ and $\{\psi_k(f_i)\}$ are the eigenvalues and eigenfunctions associated with the kernel function $\mathcal{V}(f_i, f_j)$ by Mercer's theorem (13). We achieve an orthogonal decomposition of the probability density function as

$$f(x) = \sum_{k=1}^{\infty} \psi_k(f) \xi_k(x) \quad \forall f(x) \in \mathcal{E}. \tag{20}$$

The normality condition is fulfilled by (19).

Note that the congruence map Ψ can be characterized as the unique mapping from $\mathcal{H}_{\mathcal{V}}$ into $L_2(\mathcal{E})$ satisfying the condition that for every functional $\mathcal{G}$ in $\mathcal{H}_{\mathcal{V}}$ and every j in $\mathbb{I}$

$$\int \Psi(\mathcal{G}) f_j(x)\mathrm{d}x = \langle \mathcal{G}, \mathcal{V}(f_j,\cdot)\rangle_{\mathcal{H}_{\mathcal{V}}} = \mathcal{G}(f_j). \tag{21}$$

It is obvious that Ψ in (18) fulfills (21). Then the congruence map can be represented explicitly as

$$\Psi(\mathcal{G}) = \sum_{k=1}^{\infty} a_k \xi_k(x) \quad \forall\, \mathcal{G} \in \mathcal{H}_{\mathcal{V}} \tag{22}$$

where a_k satisfies condition (15).

To prove the representation (22) is a valid and unique map, substituting (20) and (22) into (21), we obtain

$$\begin{aligned}\int \sum_{k=1}^{\infty} a_k \xi_k(x) \sum_{m=1}^{\infty} \psi_m(f_j)\xi_m(x)\mathrm{d}x &= \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} a_k \psi_m(f_j) \int \xi_k(x)\xi_m(x)\mathrm{d}x \\ &= \sum_{k=1}^{\infty} \lambda_k a_k \psi_k(f_j) = \mathcal{G}(f_j).\end{aligned}$$

In summary, we provide an explicit representation for the congruence map Ψ from RKHS $\mathcal{H}_{\mathcal{V}}$ into $L_2(\mathcal{E})$. These two spaces are equivalent in this geometrical sense. However, it should be emphasized that the constituting elements are very different in nature. The RKHS isometry framework offers a natural link between stochastic and deterministic functional analysis. Hence, it is more appealing to use RKHS $\mathcal{H}_{\mathcal{V}}$ for information-theoretic learning, as we will show in the next section.

D. Extension to Multidimensional pdfs

Extension of $\mathcal{H}_{\mathcal{V}}$ to multidimensional pdfs is straightforward since the definitions and derivations in the previous section can be easily adapted into multidimensional probability density functions. Now let $\mathcal{E}_m$ be the set of all square integrable m-dimensional probability density functions, i.e., $f_{i,m}(x_1,\ldots,x_m) \in \mathcal{E}_m$, $\forall i \in \mathbb{I}$ and $m \in \mathbb{N}$, where $\int f_{i,m}(x_1,\ldots,x_m)^2 \mathrm{d}x_1 \cdots \mathrm{d}x_m < \infty$ and $\mathbb{I}$ is the index set. We need to change the definition of kernel function (12) to

$$\mathcal{V}(f_{i,m}, f_{j,m}) = \int f_{i,m}(x_1,\ldots,x_m) \times f_{j,m}(x_1,\ldots,x_m)\mathrm{d}x_1\cdots\mathrm{d}x_m \quad \forall i,j \in \mathbb{I}.$$

Then every definition and derivation might as well be modified accordingly in the previous section. Let $\mathcal{H}_{\mathcal{V}(m)}$ denote the RKHS determined by the kernel function for m-dimensional pdfs. The proposed RKHS framework is consistent with the dimensionality of the pdfs.

The CIP based on the multidimensional pdfs characterizes the information among different random variables whose domains might not necessarily be the same in the whole space. In particular, the two-dimensional pdf CIP can be used to quantify the divergence or the cross-covariance between two random variables because the joint pdf can be factorized into a product of two marginal pdfs as a special independent case. This is exactly what the definitions of Euclidean quadratic mutual information (7) and Cauchy–Schwarz quadratic mutual information (8) are based on. We will use the two-dimensional pdf CIP to reformulate these two quantities in the following section.

IV. ITL Descriptors in the ITL RKHS Framework

In this section, we will elucidate the added insight that the ITL RKHS $\mathcal{H}_{\mathcal{V}}$ brought into the picture of ITL, the Parzen RKHS. We will also reexamine the ITL descriptors introduced in Section II.

First, as the kernel function $\mathcal{V}(f_i, f_j)$ in $\mathcal{H}_{\mathcal{V}}$ is defined as the CIP between two pdfs, immediately we have

$$\int f(x)g(x)\mathrm{d}x = \langle \mathcal{V}(f,\cdot), \mathcal{V}(g,\cdot)\rangle_{\mathcal{H}_{\mathcal{V}}}. \tag{23}$$

That is, the CIP is the inner product between two transformed functionals in the RKHS $\mathcal{H}_{\mathcal{V}}$. The inner product quantifies the similarity between two functionals, which is consistent with the definition of CIP. The IP (first moment of the pdf) can thus be specified as the inner product of the functional with respect to itself

$$\int f(x)^2 \mathrm{d}x = \langle \mathcal{V}(f,\cdot), \mathcal{V}(f,\cdot)\rangle_{\mathcal{H}_{\mathcal{V}}} = \|\mathcal{V}(f,\cdot)\|^2_{\mathcal{H}_{\mathcal{V}}}. \tag{24}$$

The IP appears as the norm square of nonlinearly transformed functional in $\mathcal{H}_{\mathcal{V}}$. Therefore, minimizing the error entropy in ITL turns out to be the maximization of the norm square in $\mathcal{H}_{\mathcal{V}}$, as seen in (1). One can expect that the norm maximization will include the information regarding the pdf. This has been recognized in ITL [27] by applying the Taylor series expansion for the Gaussian kernel used in the IP estimation (4)

$$\hat{V}(\mathrm{X}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{k=0}^{\infty}\frac{1}{k!}\left[-\frac{(x_i - x_j)^2}{2\sigma^2}\right]^k.$$

But notice that this result depends on the kernel (the Gaussian kernel just provides sums of even-order moments, a polynomial kernel will create a finite sum of moments, etc.), while now we have a clean statement that derives from the use of the first moment of the pdf in $\mathcal{H}_{\mathcal{V}}$. Moreover, the nonlinearly transformed functional $\mathcal{V}(f,\cdot)$ is deterministic in $\mathcal{H}_{\mathcal{V}}$. Hence, the proposed RKHS framework provides a link between a stochastic and a deterministic transformation.

The conventional mean square error has also been rewritten as the norm square of projected vectors in the RKHS $\mathcal{H}_R$ induced by the covariance function [3]. But the RKHS $\mathcal{H}_R$ is restricted to second-order statistics, i.e., the mean square error that is optimal for Gaussian processes, while the RKHS $\mathcal{H}_{\mathcal{V}}$ embeds the higher order statistics. Finally, $\mathcal{H}_R$ works with statistical information in the joint space of different lags, while $\mathcal{H}_{\mathcal{V}}$ would be the product of marginals at two lags. In order to access the joint space, another ITL function called *correntropy* is necessary [37], but it will not be covered in this paper.

Compared to the RKHS $\mathcal{H}_\kappa$ induced by the predesigned kernel function used in the machine learning, our framework is more elegant because it corresponds to the direct definition of the RKHS based on the first moment of the pdf without ever talking about a kernel-based pdf estimator. However, there is a close relationship between $\mathcal{H}_\mathcal{V}$ and statistics estimated in $\mathcal{H}_\kappa$, as will be fully addressed in Section VI.

Based on the reformulations of cross-information potential (23) and information potential (24) in RKHS $\mathcal{H}_\mathcal{V}$, we are ready to rewrite the one-dimensional Euclidean (5) and Cauchy–Schwarz divergence measures (6) in terms of operations on functionals in $\mathcal{H}_\mathcal{V}$. First

$$D_{\mathrm{ED}}(f,g) = \|\mathcal{V}(f,\cdot) - \mathcal{V}(g,\cdot)\|^2. \tag{25}$$

That is, the Euclidean divergence measure is in fact the norm square of the difference between two corresponding functionals in $\mathcal{H}_\mathcal{V}$, which is much more satisfying than the original description (5). The Cauchy–Schwarz divergence measure can be presented as

$$D_{\mathrm{CS}}(f,g) = -\log \frac{\langle \mathcal{V}(f,\cdot), \mathcal{V}(g,\cdot)\rangle_{\mathcal{H}_\mathcal{V}}}{\|\mathcal{V}(f,\cdot)\|\,\|\mathcal{V}(g,\cdot)\|} = -\log(\cos\theta)$$

where θ is the angle between two functional vectors $\mathcal{V}(f,\cdot)$ and $\mathcal{V}(g,\cdot)$. Therefore, the Cauchy–Schwarz divergence measure truly depicts the separation of two functional vectors in the RKHS $\mathcal{H}_\mathcal{V}$. When two vectors lie in the same direction and the angle $\theta = 0^\circ$, $D_{\mathrm{CS}}(f,g) = 0$. If two vectors are perpendicular to each other ($\theta = 90^\circ$), $D_{\mathrm{CS}}(f,g) = \infty$. The RKHS $\mathcal{H}_\mathcal{V}$ supplies rich geometric insights into the original definitions of the two divergence measures. The same conclusion has been reached by Jenssen [29] when he used the kernel estimators of $D_{\mathrm{CS}}(f,g)$ and $D_{\mathrm{ED}}(f,g)$, which shows again the close relationship between ITL RKHS and statistics evaluated in $\mathcal{H}_\kappa$.

To extend the same formulation to the Euclidean and Cauchy–Schwarz quadratic mutual information (7) and (8), consider the product of marginal pdfs $f_1(x_1)f_2(x_2)$ as a special subset $\mathcal{A}_2$ of the two-dimensional square integrable pdfs set $\mathcal{E}_2$, where the joint pdf can be factorized into product of marginals, i.e., $\mathcal{A}_2 \subseteq \mathcal{E}_2$. Then both measures characterize different geometric information between the joint pdf and the factorized marginal pdfs. The Euclidean quadratic mutual information (7) can be expressed as

$$I_{\mathrm{ED}}(\mathrm{X}_1,\mathrm{X}_2) = \|\mathcal{V}(f_{1,2},\cdot) - \mathcal{V}(f_1 f_2,\cdot)\|^2$$

where $\mathcal{V}(f_{1,2},\cdot)$ is the functional in $\mathcal{H}_{\mathcal{V}(2)}$ corresponding to the joint pdf $f_{1,2}(x_1,x_2)$ and $\mathcal{V}(f_1f_2,\cdot)$ is for the product of the marginal pdfs $f_1(x_1)f_2(x_2)$. Similarly, the Cauchy–Schwarz quadratic mutual information can be rewritten as

$$I_{\mathrm{CS}}(\mathrm{X}_1,\mathrm{X}_2) = -\log\frac{\langle \mathcal{V}(f_{1,2},\cdot), \mathcal{V}(f_1f_2,\cdot)\rangle_{\mathcal{H}_\mathcal{V}}}{\|\mathcal{V}(f_{1,2},\cdot)\|\,\|\mathcal{V}(f_1f_2,\cdot)\|} = -\log(\cos\gamma). \tag{26}$$

The angle γ is the separation between two functional vectors in $\mathcal{H}_{\mathcal{V}(2)}$. When two random variables are independent ($f_{1,2}(x_1,x_2) = f_1(x_1)f_2(x_2)$ and $\mathcal{A}_2 = \mathcal{E}_2$), $\gamma = 0^\circ$ and the divergence measure $I_{\mathrm{CS}}(\mathrm{X}_1,\mathrm{X}_2) = 0$ since two sets are equal. If $\gamma = 90^\circ$, two vectors in $\mathcal{H}_{\mathcal{V}(2)}$ are orthogonal and the joint pdf is singular to the product of marginals. In this case, the divergence measure is infinity.

The proposed RKHS framework provides an elegant and insightful geometric perspective towards information-theoretic learning. All the ITL descriptors can now be reexpressed in terms of algebraic operations on functionals in RKHS $\mathcal{H}_\mathcal{V}$. The proposed RKHS is based on the well-behaved square-integrable pdfs; therefore it excludes nonsquare-integrable pdfs. But since all the cost functions in ITL are based on square-integrable pdfs, the proposed RKHS framework is suitable for ITL and for most of statistical learning.

V. A Lower Bound for Information Potential

Based on the proposed RKHS framework for the information-theoretic learning, we derive a lower bound for the IP (2). First we cite the projection theorem in Hilbert space that we will use in the following proof.

Theorem 2 (Projection in Hilbert Space): Let $\mathcal{H}$ be a Hilbert space, $\mathcal{M}$ be a Hilbert subspace of $\mathcal{H}$ spanned by N linearly independent vectors $u_1, u_2, \ldots, u_N$, s be a vector in $\mathcal{H}$, and d be a quantity such that

$$d = \inf\|s-u\| \quad \forall u \in \mathcal{M}.$$

Then there exists a unique vector, denoted as $\mathrm{P}(s|\mathcal{M})$, in $\mathcal{M}$ such that

$$\mathrm{P}(s|\mathcal{M}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\langle s,u_i\rangle K^{-1}(i,j)u_j \tag{27}$$

where $K(i,j)$ is the $N\times N$ Gram matrix whose (i,j) is given by $\langle u_i,u_j\rangle$. The projected vector $\mathrm{P}(s|\mathcal{M})$ also satisfies the following conditions:

$$\|s-\mathrm{P}(s|\mathcal{M})\| = d = \min\|s-u_i\| \tag{28}$$

$$\langle s-\mathrm{P}(s|\mathcal{M}),u_i\rangle = 0 \quad \forall u_i\in\mathcal{M}$$

$$\langle \mathrm{P}(s|\mathcal{M}),u_i\rangle = \langle s,u_i\rangle \quad \forall u_i\in\mathcal{M}. \tag{29}$$

The geometrical explanation of the theorem is straightforward. Readers can refer to [38] for a thorough proof. Now we state the proposition on a lower bound for the information potential as a statistical estimator.

Proposition (Lower Bound for the Information Potential): Let $\mathcal{V}(f,\cdot)$ be a vector in the RKHS $\mathcal{H}_\mathcal{V}$ induced by the kernel $\mathcal{V}$ and $\mathcal{M}$ be a subspace of $\mathcal{H}_\mathcal{V}$ spanned by N linearly independent vectors $\mathcal{V}(g_1,\cdot),\mathcal{V}(g_2,\cdot),\ldots,\mathcal{V}(g_N,\cdot)\in\mathcal{H}_\mathcal{V}$. Then

$$\begin{aligned}\int f(x)^2\mathrm{d}x &= \|\mathcal{V}(f,\cdot)\|^2 \\ &\geq \sum_{i,j=1}^{N}\langle \mathcal{V}(f,\cdot),\mathcal{V}(g_i,\cdot)\rangle_{\mathcal{H}_\mathcal{V}} G^{-1}(i,j) \\ &\quad\times\langle \mathcal{V}(f,\cdot),\mathcal{V}(g_j,\cdot)\rangle_{\mathcal{H}_\mathcal{V}}\end{aligned} \tag{30}$$

where $G(i,j)$ is the $N\times N$ Gram matrix whose (i,j) term is defined as $\langle \mathcal{V}(g_i,\cdot),\mathcal{V}(g_j,\cdot)\rangle_{\mathcal{H}_\mathcal{V}}$.

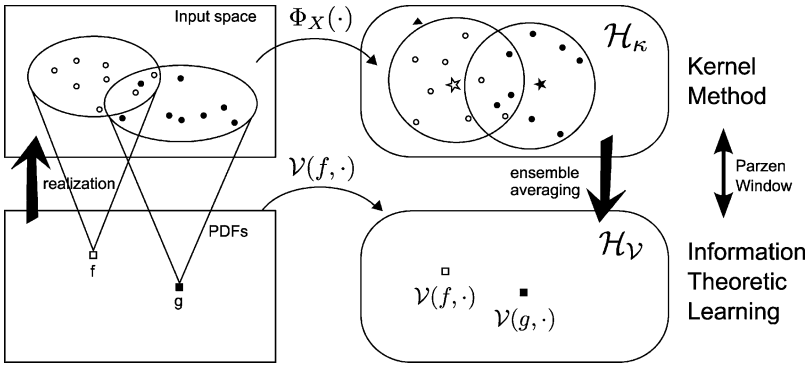


Fig. 1. The relationship among the sample space, pdf space, proposed ITL RKHS $\mathcal{H}_\mathcal{V}$, and RKHS $\mathcal{H}_\kappa$. The sample space and $\mathcal{H}_\kappa$ are connected via the nonlinear transformation $\Phi_X(\cdot)$. The pdf space and $\mathcal{H}_\mathcal{V}$ are connected via the feature map $\mathcal{V}(f,\cdot)$. A realization of a pdf in pdf space corresponds to a set of points in the sample space. The ensemble average of functionals in $\mathcal{H}_\kappa$ corresponds to one functional in $\mathcal{H}_\mathcal{V}$. The kernel methods and ITL are related via the Parzen window.

Proof: By the projection theorem (27), we can find the orthogonal projection of $\mathcal{V}(f,\cdot)$ onto the subspace $\mathcal{M}$ as

$$\mathrm{P}\left(\mathcal{V}(f,\cdot)|\mathcal{M}\right) = \sum_{i,j=1}^{N} \langle \mathcal{V}(f,\cdot), \mathcal{V}(g_i,\cdot)\rangle_{\mathcal{H}_\mathcal{V}} G^{-1}(i,j)\mathcal{V}(g_j,\cdot).$$

Since the Gram matrix is positive definite for linear independent vectors, the inverse always exists. Next, we calculate the norm square of the projected vector by $\mathcal{H}_R$ (29)

$$\begin{aligned}\left\|\mathrm{P}\left(\mathcal{V}(f,\cdot)|\mathcal{M}\right)\right\|^2 &= \langle \mathcal{V}(f,\cdot), \mathrm{P}\left(\mathcal{V}(f,\cdot)|\mathcal{M}\right)\rangle_{\mathcal{H}_\mathcal{V}} \\ &= \sum_{i,j=1}^{N} \langle \mathcal{V}(f,\cdot), \mathcal{V}(g_i,\cdot)\rangle_{\mathcal{H}_\mathcal{V}} G^{-1}(i,j) \\ &\quad \times \langle \mathcal{V}(f,\cdot), \mathcal{V}(g_j,\cdot)\rangle_{\mathcal{H}_\mathcal{V}}. \end{aligned} \tag{31}$$

On the other hand, the projection residual defined in (28) satisfies

$$d^2 = \|\mathcal{V}(f,\cdot)\|^2 - \left\|\mathrm{P}\left(\mathcal{V}(f,\cdot)|\mathcal{M}\right)\right\|^2 \geq 0. \tag{32}$$

Combining (31) and (32), we come to the conclusion of our proposition (32). □

The proposition generalizes the Cramér–Rao inequality in statistical estimation theory that only involves the variance of the estimator. It can also be viewed as an approximation to the functional norm by a set of orthogonal bases. Equation (30) offers a theoretical role for the IP in statistics.

VI. Connection Between ITL and Kernel Methods Via RKHS $\mathcal{H}_\mathcal{V}$

In this section, we connect ITL and kernel methods via the RKHS framework. As we have mentioned in the previous section, because the RKHS $\mathcal{H}_\kappa$ is induced by the data-independent kernel function, the nonlinearly projected data in $\mathcal{H}_\kappa$ are still stochastic, and statistical inference is required in order to compute quantities of interest. For instance, in order to compute the statistics of the functionals, the mean and covariance are required. The expected value of functionals in the RKHS $\mathcal{H}_\kappa$ is defined as $E[\Phi(x)]$. The cross-covariance is defined as a unique operator Σ_{XY} such that for any functionals f and g in $\mathcal{H}_\kappa$

$$\begin{aligned}\langle g, \Sigma_{\mathrm{XY}} f\rangle_{\mathcal{H}_\kappa} &= E\left[g(y)f(x)\right] - E\left[g(y)\right]E\left[f(x)\right] \\ &= \mathrm{Cov}\left[f(x), g(y)\right].\end{aligned}$$

The mean and cross-covariance operators as statistics of functionals in $\mathcal{H}_\kappa$ become intermediate steps to develop algorithms such as the maximum mean discrepancy (MMD) [39], kernel independent component analysis (kernel ICA) [20], and others. However, the proposed ITL RKHS $\mathcal{H}_\mathcal{V}$ is based on the CIP (integral of product of pdfs); therefore the transformed functional in $\mathcal{H}_\mathcal{V}$ is deterministic and only algebra is needed to carry out statistical inference in ITL RKHS. Hence our proposed RKHS offers simplicity and elegance in dealing with data statistics.

The RKHS $\mathcal{H}_\kappa$ and the RKHS $\mathcal{H}_\mathcal{V}$ are related via the expectation operator. In order to justify this statement, Parzen's nonparametric asymptotically unbiased and consistent pdf estimator (3) is employed to estimate those pdfs used in the ITL descriptors [29]. The Parzen window evaluates the pdfs in the sample space. Provided one chooses a nonnegative definite kernel function as the Parzen window, it connects the RKHS $\mathcal{H}_\mathcal{V}$ to the RKHS $\mathcal{H}_\kappa$ used in the kernel methods. As illustrated in Fig. 1, the feature map $\Phi(x)$ nonlinearly projects the sample space into a stochastic RKHS $\mathcal{H}_\kappa$. Alternatively, the feature map $\mathcal{V}(f,\cdot)$ transforms the pdf space into a deterministic RKHS $\mathcal{H}_\mathcal{V}$. Hence the stochasticity is implicitly embedded into the feature map, and immediate algebraic operation can be applied to compute statistics. However, the $\mathcal{H}_\kappa$ methodology has to rely on intermediate steps by defining mean and covariance operators.

Next we examine two kernel-based statistical methods, MMD [39] and kernel ICA [20], from the information-theoretic learning perspective. We show here that MMD is equivalent to the Euclidean divergence measure and that kernel ICA is equivalent to the Cauchy–Schwarz quadratic mutual information. The statistical computations in the RKHS $\mathcal{H}_\kappa$ have corresponding algebraic expressions in the RKHS $\mathcal{H}_\mathcal{V}$. However, note that not all kernel method algorithms can be interpreted in the ITL-RKHS framework.

A. ITL Perspective of Maximum Mean Discrepancy

The MMD [39] is a statistical test based on kernel methods to determine whether two samples are from different distributions. Theoretically, if the expected value of $p(x)$ for an arbitrary measurable function is the same for both random variables, the two distributions are identical. Since it is not practical to work with such a rich function class, MMD restricts the function class to a unit ball in an RKHS $\mathcal{H}_\kappa$ that is associated with a kernel $\kappa(\cdot,\cdot)$. This leads to the following quantity:

$$M(\mathrm{X},\mathrm{Y}) = \sup_{\|p\|_{\mathcal{H}_\kappa}\le 1} \left(E\left[p(X)\right] - E\left[p(Y)\right]\right) \tag{33}$$

where X and Y are the underlying random variables of the two distributions and p is the family of measurable functionals in the unit ball of the RKHS $\mathcal{H}_\kappa$.

The kernel trick can be employed here to compute MMD, that is

$$p(x) = \langle \Phi(x), p\rangle_{\mathcal{H}_\kappa} = \langle \kappa(x,\cdot), p\rangle_{\mathcal{H}_\kappa}. \tag{34}$$

Substituting (34) into the definition of MMD (33), we obtain

$$M(\mathrm{X},\mathrm{Y}) = \|m_{\mathrm{X}} - m_{\mathrm{Y}}\|_{\mathcal{H}_\kappa}$$

where $m_{\mathrm{X}} = E[\Phi(X)]$ and $m_{\mathrm{Y}} = E[\Phi(Y)]$ are the statistical expectations of the functionals $\Phi(x)$ and $\Phi(y)$ in $\mathcal{H}_\kappa$. Applying $m_{\mathrm{X}} = (1/N)\sum_{i=1}^{N}\Phi(x_i)$ and $\kappa(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}_\kappa}$, an empirical estimate of MMD can be obtained as

$$\left|\hat{M}(\mathrm{X},\mathrm{Y})\right|^2 = \frac{1}{N^2}\sum_{i,j=1}^{N}\kappa(x_i,x_j) - \frac{2}{NL}\sum_{i,j=1}^{N,L}\kappa(x_i,y_j) + \frac{1}{L^2}\sum_{i,j=1}^{L}\kappa(y_i,y_j) \tag{35}$$

where $\{x_i\}_{i=1}^{N}$ and $\{y_j\}_{j=1}^{L}$ are two sets of data samples. The estimate of MMD provides a statistical test to determine whether two sets of data samples are from the same distribution.

We now prove that MMD is equivalent to the Euclidean divergence measure. To transform the Euclidean divergence measure (5) to the sample space, we use the Parzen window (3) to estimate the pdfs given two sets of samples $\{x_i\}_{i=1}^{N}$ and $\{y_j\}_{j=1}^{L}$ to obtain an empirical value as

$$\hat{D}_{\mathrm{ED}}(f,g) = \frac{1}{N^2}\sum_{i,j=1}^{N}\kappa(x_i,x_j) - \frac{2}{NL}\sum_{i,j=1}^{N,L}\kappa(x_i,y_j) + \frac{1}{L^2}\sum_{i,j=1}^{L}\kappa(y_i,y_j). \tag{36}$$

Comparing (35) and (36), we observe that $|\hat{M}(\mathrm{X},\mathrm{Y})|^2 = \hat{D}_{ED}(f,g)$. Since the Euclidean divergence measure can be rewritten as the norm square of difference between two functionals in the RKHS $\mathcal{H}_\mathcal{V}$ (25), we obtain

$$\|m_{\mathrm{X}} - m_{\mathrm{Y}}\|^2_{\mathcal{H}_\kappa} = \|\mathcal{V}(f,\cdot) - \mathcal{V}(g,\cdot)\|^2_{\mathcal{H}_\mathcal{V}}. \tag{37}$$

The left-hand side is the norm square of difference between two functional expectations in the RKHS $\mathcal{H}_\kappa$. Since the functional $\Phi(x)$ is still stochastic in $\mathcal{H}_\kappa$, the expectation operation is necessary to carry out the computation. On the other hand, the right-hand side is the norm square of difference between two functionals in the RKHS $\mathcal{H}_\mathcal{V}$ (25). Because the functional $\mathcal{V}(f,\cdot)$ is deterministic, the computation is algebraic. The feature map $\mathcal{V}(f,\cdot)$ for the RKHS $\mathcal{H}_\mathcal{V}$ is equivalent to the expectation of the feature map $\Phi(x)$ for the RKHS $\mathcal{H}_\kappa$. Therefore, the proposed RKHS framework provides a natural link between stochastic and deterministic functional analysis. The MMD in kernel methods is essentially equivalent to the Euclidean divergence measure in ITL.

B. ITL Perspective of Kernel ICA

Kernel ICA is a novel independent component analysis method based on a kernel measure of dependence [20]. It assumes an RKHS $\mathcal{H}_\kappa$ determined by the kernel $\kappa(x,y)$ and the feature map $\Phi(x)$. The feature map $\Phi(x)$ can be derived from the eigendecomposition of the kernel function $\kappa(x,y)$ according to Mercer's theorem and forms an orthogonal basis for the RKHS $\mathcal{H}_\kappa$. Then the $\mathcal{F}$-correlation function is defined as the maximal correlation between the two random variables $f_1(\mathrm{X}_1)$ and $f_2(\mathrm{X}_2)$, where f_1 and f_2 range over $\mathcal{H}_\kappa$

$$\begin{aligned}\rho &= \max_{f_1,f_2}\operatorname{corr}\left(f_1(X_1), f_2(X_2)\right)\\ &= \max_{f_1,f_2}\frac{\operatorname{cov}\left(f_1(X_1), f_2(X_2)\right)}{\sqrt{\operatorname{var}\left(f_1(X_1)\right)\operatorname{var}\left(f_2(X_2)\right)}}.\end{aligned} \tag{38}$$

Obviously, if the random variables X_1 and X_2 are independent, then the $\mathcal{F}$-correlation is zero. The converse is also true provided that the RKHS $\mathcal{H}_\kappa$ is large enough. This means that $\rho = 0$ implies X_1 and X_2 are independent.

In order to obtain a computationally tractable implementation of $\mathcal{F}$-correlation, the reproducing property of RKHS (i.e., kernel trick) (34) is used to estimate the $\mathcal{F}$-correlation. The nonlinear functionals f_1 and f_2 can be represented by the linear combination of the basis $\{\Phi(x^i)\}_{i=1}^{N}$ in which $\{x^i\}_{i=1}^{N}$ is N empirical observations of random variable X. That is

$$f_1 = \sum_{k=1}^{N}\alpha_1^k\Phi\left(x_1^k\right), \quad \text{and} \quad f_2 = \sum_{k=1}^{N}\alpha_2^k\Phi\left(x_2^k\right). \tag{39}$$

Substituting (39) and (34) into (38) and using the empirical data to approximate the population value, the $\mathcal{F}$-correlation can be estimated as

$$\hat{\rho} = \max_{\alpha_1,\alpha_2\in\mathbb{R}^N}\frac{\alpha_1^\top K_1 K_2 \alpha_2}{\sqrt{\left(\alpha_1^\top K_1^2\alpha_1\right)\left(\alpha_2^\top K_2^2\alpha_2\right)}} \tag{40}$$

where K_1 and K_2 are the Gram matrices associated with the data sets $\{x_1^i\}_{i=1}^{N}$ and $\{x_2^i\}_{i=1}^{N}$ defined as $[K_l]_{i,j} = \kappa(x_l^i, x_l^j)$.

Because the cost function in (40) is not a numerically stable estimator in general, regularization is needed by penalizing the RKHS norms of f_1 and f_2 in the denominator of (40). The regularized estimator has the same independence characterization property of the $\mathcal{F}$-correlation as (40), since it is the numerator

$\alpha_1^\top K_1 K_2 \alpha_2$ in the $\mathcal{F}$-correlation that characterizes the independence property of two random variables. The difference between the direct estimator (40) and the regularized version is only the normalization.

We prove here the equivalence between the cost function used in kernel ICA (40) and the Cauchy–Schwarz quadratic mutual information (8). To prove the equivalence, we use the weighted Parzen window, which is defined as

$$\hat{f}(x) = \frac{1}{A}\sum_{i=1}^{N} \alpha_i \kappa(x, x_i) \tag{41}$$

where A is a normalization term such that the integral of $\hat{f}(x)$ equals one.

When the Cauchy–Schwarz quadratic mutual information (8) is used as a contrast function in ICA, it should be minimized so that the mutual information between random variables is also minimized. As the logarithm is a monotonic function, minimizing the Cauchy–Schwarz quadratic mutual information is equivalent to maximizing its argument. Therefore, by approximating the population expectation with sample mean for the argument in (8) and estimating the joint and marginal pdfs with weighed Parzen window (41), we obtain

$$\hat{J} = \max_{\alpha_1, \alpha_2 \in \mathbb{R}^N} \frac{\alpha_1^\top K_1 K_2 \alpha_2}{\sqrt{L(\mathbf{1}^\top K_1 \alpha_1)(\mathbf{1}^\top K_2 \alpha_2)}} \tag{42}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^\top$, $[K_l]_{i,j} = \kappa(x_l^i, x_l^j)$, and $L = \sum_{i,j}^{N} \alpha_1^i \kappa(x_1^i, x_1^j)\kappa(x_2^i, x_2^j)\alpha_2^i$.

Comparing (40) and (42), we notice that they have the same numerators and different normalizations. As we already pointed out, it is the numerators in the kernel ICA and the Cauchy–Schwarz quadratic mutual information that characterize the dependence measure of two random variables. The denominators only provide normalization. Hence we conclude that the Cauchy–Schwarz quadratic mutual information, estimated via weighed Parzen window, is equivalent to the kernel ICA. Moreover, the coordinates of the nonlinear functionals f_1 and f_2 in the RKHS $\mathcal{H}_\kappa$ (39) have corresponding terms in the weighted Parzen window (41).

In summary, the feature map $\Phi(x)$ works with individual data samples and transforms each datum into the RKHS $\mathcal{H}_\kappa$ induced by the kernel $\kappa(\cdot,\cdot)$. For applications involving statistical inference on the transformed data, extra operators such as the mean and covariance are required. On the other hand, the feature map $\mathcal{V}(f,\cdot)$ deals with the pdf directly and transforms each pdf into the RKHS $\mathcal{H}_\mathcal{V}$ determined by the kernel $\mathcal{V}(\cdot,\cdot)$. If the applications are based on the statistics of the transformed functionals, only algebraic computation is needed without defining any extra operators as required in RKHS $\mathcal{H}_\kappa$. Therefore the proposed RKHS framework provides a direct and elegant treatment of statistical inference using the RKHS technique. Certainly, the RKHS $\mathcal{H}_\kappa$ is more flexible in other applications beyond statistical inference since it is based on the available data samples. The RKHS $\mathcal{H}_\mathcal{V}$ is built directly upon pdfs and requires Parzen window to evaluate the overall cost functions from samples.

VII. Discussions

In this section, we relate our work to the concepts of information geometry and probability product kernels.

A. Nonparametric Versus Parametric Modeling

The RKHS framework presented in this paper elucidates the geometric structure on the space of finite square integrable probability density functions. Since no model for the pdf is assumed, it is nonparametric and infinite-dimensional. In statistics, information geometry studies the intrinsic geometry in a finite-dimensional parametric statistical manifold formed by the pdfs [40]. Extension to infinite-dimensional nonparametric submanifold has been made [41]. For finite-dimensional parametric families of pdfs, the only invariant metric to the tangent space is the Riemannian structure defined by the Fisher information [40], [42]

$$g_{ij}(\theta) = \mathbf{E}\left[\frac{\partial \log f(x;\theta)}{\partial \theta_i}\frac{\partial \log f(x;\theta)}{\partial \theta_j}\right]$$

in the component form. The Riemannian metric coincides locally (infinitesimally) with the double of the Kullback–Leibler divergence. More interestingly, the Fisher information is a symmetric nonnegative definite function defined in the parameter space. Therefore, it also uniquely determines a reproducing kernel Hilbert space. But, it is very different from our approach because the ITL RKHS framework is model free, i.e., it does not assume a parametric family of probability density functions. The nonnegative definite kernel function (12) is defined directly in the pdf space; however, the kernel function in information geometry is defined in parameter space since it aims at estimating model parameters from data. Hence, both methodologies define nonparametric and parametric reproducing kernel Hilbert spaces, respectively, to tackle problems of interest from different perspectives.

B. Kernel Function as a Similarity Measure

The kernel function we defined characterizes relationships between probability density functions. For instance, the one-dimensional kernel function $\mathcal{V}(f,g)$ quantifies how similar one pdf is to the other. The two-dimensional function $\mathcal{V}(f_{1,2}, f_1 f_2)$ specifies the relationship between the joint pdf and the product of marginal pdfs. Therefore it also measures how dependent one random variable is on the other. The Cauchy–Schwarz quadratic mutual information (26) was applied to independent component analysis based on this interpretation [33]. Using probability distributions to measure similarity is nothing new. One customary quantity is the Kullback–Leibler divergence. However, it is not positive definite nor symmetric, and hence does not have an RKHS associated with it. Therefore, Kullback–Leibler divergence lacks the geometric advantage of RKHS that our kernel function possesses.

Recently, several probability product kernels have been proposed in machine learning to use ensembles instead of individual data to capture dependence between generative models [43]. It is shown that the Bhattacharyya coefficient defined by

$$\kappa(f_i, f_j) = \int \sqrt{f_i(x) f_j(x)}\mathrm{d}x$$

is a reproducing kernel [44]. The expected likelihood kernel in [44] is exactly the CIP, but since it was proposed purely from a machine learning point of view, it failed to elucidate its broader information-theoretic underpinning. With our approach to propose an RKHS framework for information-theoretic learning, construct the RKHS bottom-up, and prove its validity mathematically, the kernel function embodies a rich information-theoretic interpretation. Moreover, as ITL is mainly applied in adaptive signal processing, a nonparametric method is employed to compute the CIP kernel and other quantities without an explicit probability density function estimation. However, previous approaches assume a parametric generative model in order to calculate the kernel in their approach [43], [44].

A family of Hilbertian metrics has also been proposed recently for probability measures induced from nonnegative kernels [45]. The divergence measures we presented are related to some of the members; however, the space of probability measures is not explicitly manipulated as an RKHS. Thus this work would shed light on the understanding of the Hilbertian metrics. The family also includes various other possibilities for divergence measures and Hilbert spaces for the pdfs, but not all of them have efficient estimators from samples as our method provides.

VIII. Conclusion

In this paper, we present a geometric structure for the information-theoretic learning methodology. The proposed reproducing kernel Hilbert space framework is determined by the symmetric nonnegative definite kernel function, which is defined as the cross-information potential. The kernel function quantifies the similarity between two transformed functionals in the RKHS $\mathcal{H}_{\mathcal{V}}$. We can rewrite the ITL descriptors of entropy and divergence and associated cost functions in terms of algebraic operations on the functionals in the proposed RKHS. This formulation offers a solid foundation and rich geometry for the original information-theoretic learning algorithms. Compared to the previous RKHS frameworks, the proposed ITL RKHS is built directly on the probability density functions and contains the statistical information of the data. Hence, the RKHS $\mathcal{H}_{\mathcal{V}}$ provides an elegant geometric structure intrinsic to the data. We also elucidate the natural link between ITL and the kernel methods via the proposed ITL RKHS framework. Future work will include deriving least projection theorem in RKHS $\mathcal{H}_{\mathcal{V}}$ so that we might present minimum information potential estimator and others directly from the functionals. Another central issue in kernel methods is related to kernel design. The ITL RKHS may provide another avenue to specify the kernel properties required to conduct statistical inference and address the fundamental problem of dependence measures.

References

[1] E. H. Moore, "On properly positive Hermitian matrices," *Bull. Amer. Math. Soc.*, vol. 23, no. 59, pp. 66–67, 1916.

[2] N. Aronszajn, "The theory of reproducing kernels and their applications," *Cambridge Phil. Soc. Proc.*, vol. 39, pp. 133–153, 1943.

[3] E. Parzen, "Statistical inference on time series by Hilbert space methods," Statistics Dept., Stanford Univ., Stanford, CA, Tech. Rep. 23, 1959.

[4] E. Parzen, "An approach to time series analysis," *Ann. Math. Stat.*, vol. 32, no. 4, pp. 951–989, Dec. 1961.

[5] E. Parzen, "Extraction and detection problems and reproducing kernel Hilbert spaces," *SIAM J. Contr.*, vol. 1, pp. 35–62, 1962.

[6] T. Kailath, "RKHS approach to detection and estimation problems—Part I: Deterministic signals in Gaussian noise," *IEEE Trans. Inf. Theory*, vol. IT-17, pp. 530–549, Sep. 1971.

[7] T. Kailath and H. Weinert, "An RKHS approach to detection and estimation problems—Part II: Gaussian signal detection," *IEEE Trans. Inf. Theory*, vol. IT-21, pp. 15–23, Jan. 1975.

[8] T. Kailath and D. Duttweiler, "An RKHS approach to detection and estimation problems—Part III: Generalized innovations representations and a likelihood-ratio formula," *IEEE Trans. Inf. Theory*, vol. IT-18, pp. 730–745, Nov. 1972.

[9] D. Duttweiler and T. Kailath, "RKHS approach to detection and estimation problems—Part IV: Non-Gaussian detection," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 19–28, Jan. 1973.

[10] D. Duttweiler and T. Kailath, "RKHS approach to detection and estimation problems—Part V: Parameter estimation," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 29–37, Jan. 1973.

[11] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990, vol. 49.

[12] R. J. DeFigueiredo, "A generalized Fock space framework for nonlinear system and signal analysis," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 637–647, Sep. 1983.

[13] V. Fock, *The Theory of Space Time and Gravitation*. London, U.K.: Pergamon, 1959.

[14] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[15] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Disc.*, vol. 2, no. 2, pp. 121–167, 1998.

[16] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *J. Mach. Learn. Res.*, vol. 2, pp. 299–312, 2001.

[17] T. M. Cover, "Classification and generalization capabilities of linear threshold units," Rome Air Force, Italy, Tech. Rep. RADC-TDR-64-32, Feb. 1964.

[18] B. Schölkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[19] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neur. Comput.*, vol. 10, pp. 1299–1319, 1998.

[20] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2002.

[21] J. C. Principe, D. Xu, and J. W. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000, pp. 265–319.

[22] D. Erdogmus and J. C. Principe, "From linear adaptive filtering to nonlinear information processing," *IEEE Signal Process. Mag.*, vol. 23, pp. 14–33, Nov. 2006.

[23] A. Rényi, "On measures of entropy and information," in *Selected Papers of A. Rényi*. Budapest, Hungary: Akademiai Kiado, 1976, vol. 2, pp. 565–580.

[24] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 158–171, Feb. 2002.

[25] R. Jenssen and T. Eltoft, "An information theoretic perspective to Mercer kernel-based clustering," *Neurocomputing*, 2008, to be published.

[26] I. Hild, K. E. , D. Erdogmus, K. Torkkola, and J. Principe, "Feature extraction using information-theoretic learning," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, pp. 1385–1392, Sep. 2006.

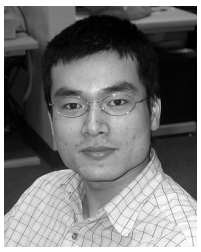
[27] D. Erdogmus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1780–1786, Jul. 2002.

[28] M. Lazaro, I. Santamaria, D. Erdogmus, K. Hild, C. Pantaleon, and J. Principe, "Stochastic blind equalization based on pdf fitting using Parzen estimator," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 696–704, Feb. 2005.

[29] R. Jenssen, D. Erdogmus, J. Principe, and T. Eltoft, "Some equivalences between kernel methods and information theoretic methods," *J. VLSI Signal Process.*, vol. 45, pp. 49–65, 2006.

[30] A. N. Kolmogorov, "Sur la notion de la moyenne," *Atti della Reale Accademia Nazionale dei Lincei, Serie VI, Rendiconti*, vol. 12, pp. 388–391, 1930.

[31] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sept. 1962.
[32] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[33] D. Xu, J. C. Principe, J. W. Fisher, and H. C. Wu, "A novel measure for independent component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1998, vol. 2, pp. 12–15.
[34] R. Jenssen, D. Erdogmus, J. C. Principe, and T. Eltoft, "The Laplacian PDF distance: A cost function for clustering in a kernel feature space," in *Advances in Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2004, pp. 625–632.
[35] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Phil. Trans. Roy. Soc. London*, vol. 209, pp. 415–446, 1909.
[36] R. Jenssen, J. C. Principe, D. Erdogmus, and T. Eltoft, "The Cauchy-Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels," *J. Franklin Inst.*, vol. 343, pp. 614–629, 2006.
[37] I. Santamaria, P. Pokharel, and J. C. Príncipe, "Generalized correlation function: Definition, properties, and application to blind equalization," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2187–2197, Jun. 2006.
[38] E. Kreyszig, *Introductory Functional Analysis with Applications*. New York: Wiley, 1978.
[39] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2006.
[40] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence, RI: AMS/Oxford Univ. Press, 2000.
[41] B. Pistone and C. Sempi, "An infinite-dimensional geometric structure on the space of all probability measures equivalent to a given one," *Ann. Statist.*, vol. 23, no. 5, pp. 1543–1561, 1995.
[42] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bull. Cal. Math. Soc.*, vol. 37, pp. 81–91, 1945.
[43] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *J. Mach. Learn. Res.*, vol. 5, pp. 819–844, 2004.
[44] T. Jebara and R. Kondor, "Bhattarcharyya and expected likelihood kernels," presented at the Conf. Learning Theory (COLT), Washington, DC, 2003.
[45] M. Hein and O. Bousquet, "Hilbertian metrics and positive definite kernels on probability measures," in *Proc. 10th Int. Workshop on Artificial Intelligence Statistics (AISTATS)*, 2005, pp. 136–143.



Jian-Wu Xu (S'03) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2002 and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 2007.

Since 2002, he has been with the Computational NeuroEngineering Laboratory, University of Florida, under the supervision of Dr. J. C. Principe. His current research interests include information theoretic learning, adaptive signal processing, control, and machine learning.

Dr. Xu is a member of Tau Beta Pi and Eta Kappa Nu.

António R. C. Paiva (S'03) was born in Ovar, Portugal, in 1980. He received the B.S. degree in electronics and telecommunications engineering from the University of Aveiro, Portugal, in 2003 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, in 2005 and 2008, respectively.

After completing his undergraduate studies, he conducted research in image compression as a research assistant of Dr. A. Pinho for almost a year. In fall 2004, he joined the Computational NeuroEngineering Laboratory, University of Florida, working under the supervision of Dr. J. C. Principe. His doctoral research focused on the development of a reproducing kernel Hilbert spaces framework for analysis and processing of point processes, with applications on single-unit neural spike trains. His research interests are, broadly, signal and image processing, with special interest in biomedical and biological applications. In particular, these include: kernel methods and information theoretic learning, image processing, brain-inspired computation, principles of information representation and processing in the brain, sensory and motor systems, and development of biological and biomedical data analysis methods.

During his undergraduate studies, Dr. Paiva received four merit scholarships, a Dr. Vale Guimaraes award for Best District Student at the University of Aveiro, and an Eng. Ferreira Pinto Basto award from Alcatel Portugal for top graduating student in the major.

Il Park (Memming) received the B.S. degree in computer science from Korea Advanced Institute of Science and Technology, South Korea, in 2004 and the M.S. degree in electrical and computer engineering from the University of Florida, Gainesville, in 2007, where he currently is pursuing the Ph.D. degree in biomedical engineering.

He has been with the Computational NeuroEngineering Laboratory, University of Florida, since 2005 under the supervision of Dr. J. C. Principe since 2005. His research interests include neural computation, signal processing, and machine learning.

Jose C. Principe (F'00) is Distinguished Professor of Electrical and Biomedical Engineering at the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANNs) modeling. He is BellSouth Professor and Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). He is involved in biomedical signal processing, in particular the electroencephalogram and the modeling and applications of adaptive systems. He is past President of the International Neural Network Society. He is a former Member of the Scientific Board of the Food and Drug Administration. He has been Supervisory Committee Chair of more than 50 Ph.D. and 61 master's students. He is author of more than 400 refereed publications.

Dr. Principe is an AIMBE Fellow. He is former Editor-in-Chief of IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and current Editor-in-Chief of IEEE REVIEWS IN BIOMEDICAL ENGINEERING. He is formal Secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. He received the IEEE Engineering in Medicine and Biology Society Career Service Award.